心理科学进展　2022, Vol. 30, No. 5, 1168−1182
Advances in Psychological Science

https://doi.org/10.3724/SP.J.1042.2022.01168

·研究方法(Research Method)·

计算机化分类测验终止规则的类别、特点及应用*

任　赫　黄颖诗　陈　平

(北京师范大学中国基础教育质量监测协同创新中心, 北京　100875)

摘　要　计算机化分类测验(Computerized Classification Testing, CCT)能够高效地对被试进行分类, 已广泛应用于合格性测验及临床心理学中。作为 CCT 的重要组成部分, 终止规则决定测验何时停止以及将被试最终划分到何种类别, 因此直接影响测验效率及分类准确率。已有的三大类终止规则(似然比规则、贝叶斯决策理论规则及置信区间规则)的核心思想分别为构造假设检验、设计损失函数和比较置信区间相对位置。同时, 在不同测验情境下, CCT 的终止规则发展出不同的具体形式。未来研究可以继续开发贝叶斯规则、考虑多维多类别情境以及结合作答时间和机器学习算法。针对测验实际需求, 三类终止规则在合格性测验上均有应用潜力, 而临床问卷则倾向应用贝叶斯规则。

关键词　计算机化分类测验, 终止规则, 似然比, 随机缩减, 贝叶斯决策理论
分类号　B841

1　引言

由于能够改变传统纸笔测验中相对固化的试题形式、更深刻地体现“因材施测”和“高效施测”, 计算机测验尤其是计算机化自适应测验(Computerized Adaptive Testing, CAT)近年来得到飞速发展。对于 CAT 而言, 其测验目的一般是准确估计被试能力, 而计算机化分类测验(Computerized Classification Testing, CCT)——作为 CAT 的一个重要分支——则以分类考生为目的。具体来说, CCT 在 CAT 的基础上可以根据预设的分界分数将被试划分到两个(比如, 掌握和未掌握)或多个(比如, 合格、良好和优秀)不同的类别中。相比于传统纸笔测验, CCT 的优势在于: 首先, CCT 不仅可以自适应地呈现最适合被试作答的题目, 还可以在保持相同决策精度的情况下大大缩短测验长度(Spray & Reckase, 1996), 进而降低测验成本、减少被试疲劳效应的影响; 其次, CCT 依托于计算机施测的特点使其能够为被试呈现更加丰富的测验内容和题目形式(比如交互式测评), 并获取更多元细致的被试数据; 再者, CCT 的高效计算力使得更精细测量模型和算法的使用成为可能, 比如融入过程性或多模态数据的模型(Sie et al., 2015; Zhan et al., 2021)以进一步满足各种测验需求、提升分类决策的可靠性。目前, CCT 已经在合格性测验(比如, 职业资格考试)以及临床心理学或医学诊断(比如, 焦虑、抑郁等精神疾病的自我报告问卷和健康与护理问卷)中得到广泛应用(Finkelman et al., 2011; Huebner & Fina, 2015; Smits & Finkelman, 2013)。

作为 CAT 的特例, 完整的 CCT 同样包括心理测量模型、标定的题库、选题策略、能力参数估计方法以及终止规则五个核心部分。但是如前所述, 两者在测验目的上并不相同: CAT 的目的是对被试能力进行准确估计(陈平, 2016), 而 CCT 是要对被试的类别进行准确划分。因此, 终止规则是区分 CCT 与 CAT 的一项主要特征(任赫, 陈平, 2021)。总体而言, CCT 终止规则关注的核心问题是系统是否有足够的把握将被试划分到某个特定的类别, 或者说系统是否可以接受当前的决策结

收稿日期: 2021-06-18
* 国家自然科学基金面上项目(32071092)、中国基础教育质量监测协同创新中心基础教育质量监测科研基金项目(2019-01-082-BZK01 和 2019-01-082-BZK02)资助。
通信作者: 陈平, E-mail: pchen@bnu.edu.cn

果(比如：继续测验、将被试划分到掌握/未掌握类别)可能产生的成本(如：测验效率的牺牲、第 I 类或第 II 类错误率)。由此, 终止规则决定测验何时停止以及将被试最终划分到何种类别, 将直接影响测验的效率和分类准确率。已有的 CCT 终止规则包括定长(fixed-length)的规则(即每名被试作答固定数量的题目)以及变长(variable-length)的规则(即每名被试作答数量不定的题目)。定长的规则比较简单, 不再赘述, 本文主要关注变长的规则。需要指出的是, 尽管定长终止规则的效率较低, 但是它可以保证所有被试作答相同长度的测验, 能够减少被试对测验公平的质疑, 主要应用于高利害测验中。与之相对应, 变长的规则具有高效的特点, 能够大大地缩短测验长度, 可以广泛应用于各类低利害测验中。

变长CCT的实施过程可以看作一种序贯抽样方案, 即“在抽样时不规定总的抽样个数, 而是根据已抽取的样本结果决定是否继续抽样, 直至停止”。最早的变长终止规则是 Ferguson (1969)根据序贯检验(Wald, 1947)提出的序贯似然比方法(Sequential Probability Ratio Test, SPRT)。SPRT 方法通过事先设定第 I 和第 II 类错误率来控制不同决策的损失, 并使用二项分布对被试作答进行建模, 相当于假设题库中所有题目的正确作答概率相同, 相应地以随机或固定顺序呈现题目。但是, Lewis 和 Sheehan (1990)则认为应该在测验过程中直接控制每一步可能造成的损失, 这就需要利用贝叶斯理论进行决策。另外, 为了使序贯抽样过程能够与被试能力相适应, Reckase (1983)与 Kingsbury 和 Weiss (1983)分别引入项目反应理论(Item Response Theory, IRT)模型。前者使用 IRT 模型代替二项分布, 进而发展出允许自适应选题的 SPRT 方法(也即对 Ferguson 方法的改进), 而后者利用能力估计的置信区间进行分类决策。综上, 前人分别从不同的视角出发, 基于不同的统计学理论建构出三类终止规则, 它们分别是似然比规则、贝叶斯决策理论规则(后文简称贝叶斯规则)和置信区间规则(Ability Confidence Intervals, ACI)。

此外, 在构造具体的 CCT 终止规则时, 还需要考虑不同测验情境的特点, 主要包括被试的类别数和测验的维度数。在被试类别方面, 有时只需要将被试划分到两个不同类别, 而有时则需要将被试划分到三个及以上的不同类别, 它们分别对应于二分类的 CCT 与多分类的 CCT。在测验维度方面, 一些测验只需要考虑被试在单个维度上的潜在特质, 但是更多的心理或教育测验往往需要同时考察被试在多个维度上的潜在特质(康春花, 辛涛, 2010), 这就分别对应于单维 CCT (Unidimensional CCT, UCCT)与多维 CCT (Multidimensional CCT, MCCT)。需要说明的是, 多分类的 CCT 终止规则在构造上与二分类的相比有较大差异, 而 MCCT 的终止规则通常可以由 UCCT 经过较为直接的推广而得到。

基于此, 本文将结合不同的测验情境, 对似然比规则、贝叶斯规则以及置信区间规则分别进行详细述评, 然后对各种规则的优劣进行讨论分析, 最后对 CCT 终止规则的未来研究方向及应用进行说明。

2　似然比规则

似然比规则的核心是通过构造似然比统计量(likelihood ratio statistics)进行假设检验。一般而言, 似然比规则的构建主要包含 4 个步骤(任赫, 陈平, 2021)：(1)构造被试能力属于特定类别的假设检验; (2)在任意两个相邻的类别之间确定能力阈值; (3)在每个阈值的上下给定一个 δ 邻域。当被试的能力值落在该区间时, 认为被试能力与阈值没有显著差异, 因此该区间也被称为无差别区间; (4)构建似然比统计量并确定拒绝域。具体地说, 根据不同的假设检验与统计量, 似然比规则可以被划分为 SPRT 和广义似然比方法(Generalized Likelihood Ratio, GLR), 下面进行详细介绍。

2.1　序贯似然比方法(SPRT)

2.1.1　二分类的 SPRT 方法

最早的二分类似然比终止规则就是 Wald (1947)提出的 SPRT。在此基础上, 研究者们主要致力于解决两个方面的问题：一是如何进一步提升二分类 SPRT 的决策效率; 二是如何将单维的二分类 SPRT 拓展到多维情境。对于第一个问题, Finkelman (2003, 2010)将随机缩减(stochastic curtailment)技术与 SPRT 方法相结合,提出随机缩减的 SPRT (Stochastically Curtailed SPRT, SCSPRT), 以进一步提高测验效率。需要指出的是, 上述方法仅适用于单维情境。对于第二个问题, 即将已有方法推广至 MCCT 时, 规则的构建思路基本没有变化, 但是能力参数的多维性会导致

UCCT 中的能力分界点转变为多维空间中的能力分界曲线或曲面(任赫, 陈平, 2021)。为此, Nydick (2013)从两个不同的角度解决这一问题, 分别提出约束的 SPRT (Constrained SPRT, C-SPRT) 以及使用空间投影方法构建的投影 SPRT (Projected SPRT, P-SPRT)。另外, Nydick (2013)还在 C-SPRT 的基础上结合随机缩减技术开发出随机缩减的多维 SPRT (Multidimensional SCSPRT, M-SCSPRT)。下文依次介绍单维的 SPRT 与 SCSPRT 以及多维的 C-SPRT、P-SPRT 与 M-SCSPRT。

(1)单维的 SPRT 方法(SPRT 与 SCSPRT)

在 UCCT 中, SPRT 使用一组简单假设来判断被试的能力分类, 即

$$H_0: \theta = \theta_l = \theta_0 - \delta$$
$$H_1: \theta = \theta_u = \theta_0 + \delta, \tag{1}$$

其中, θ 是被试的能力值, θ_0 为事先确定的分界分数, δ 为邻域大小的一半, 即无差别区间宽度的一半, $[\theta_l, \theta_u]$ 就是构造的无差别区间。由此, H_0 表示被试恰好被划入未掌握的一类, H_1 表示恰好被划入掌握的一类。

由此, SPRT (Wald, 1947)构造对数似然比统计量如下,

$$C_{ij'} = \log[\mathrm{LR}(\theta_u, \theta_l \mid \boldsymbol{Y}_{i,j'})] = \log\left[\frac{L(\theta_u \mid \boldsymbol{Y}_{i,j'})}{L(\theta_l \mid \boldsymbol{Y}_{i,j'})}\right], \tag{2}$$

其中, $L(\theta \mid \boldsymbol{Y}_{i,j'})$ 为基于 IRT 的似然函数, $\boldsymbol{Y}_{i,j'} = (Y_{i1}, Y_{i2}, \cdots, Y_{ij'})$ 为被试 i 在题目 $j = 1, 2, \cdots, j'$ 上的作答向量。记第 I 类、第 II 类错误率分别为 α 和 β, 令 $A = \beta/(1-\alpha)$ 、 $B = (1-\beta)/\alpha$ 、 $C_l = log(A)$ 、 $C_u = log(B)$ (Finkelman, 2003)。被试 i 完成 j' 道题目后, 计算对数似然比统计量 $C_{ij'}$, 并按如下规则对被试给出判断: 若 $C_{ij'} \leqslant C_l$, 则考生的分数更有可能低于分数线, 判断被试属于“未掌握”, 并结束测验, 记测验长度为 j'; 若 $C_{ij'} \geqslant C_u$, 则考生的分数更有可能高于分数线, 判断被试属于“掌握”, 并结束测验, 记测验长度为 j'; 否则, 要求被试继续作答下一道题。例如, 图 1 展示了使用两参数逻辑斯蒂克模型模拟数据得到的“不同能力取值下的对数似然函数值”, 当分界分数取 -0.5 、 $\delta = 0.1$ 、 $\alpha = \beta = 0.05$ 时, 得到 $\theta_l = -0.6$ 、 $\theta_u = -0.4$ 、 $C_l = -2.94$ 、 $C_u = 2.94$ 。此时, $\log[\mathrm{L}(\theta_u \mid \boldsymbol{Y}_{i,j'})] = -9.68, \log[\mathrm{L}(\theta_l \mid \boldsymbol{Y}_{i,j'})] = -12.41$, 于是计算得到对数似然比统计量 $C_{ij'} = (-9.68) - (-12.41) = 2.73$ 。由于 $C_l < C_{ij'} < C_u$, 所以继续测验。

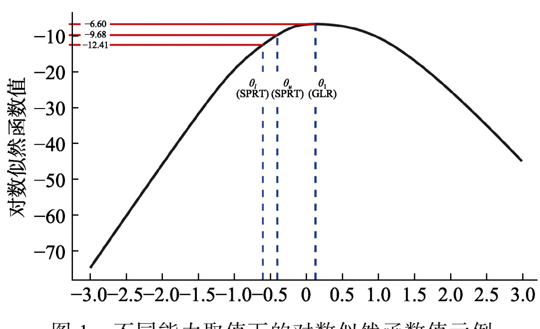


图 1 不同能力取值下的对数似然函数值示例

Wald-Wolfowitz 定理(Wald & Wolfowitz, 1948)表明: 在“测验持续进行直至满足 $C_{ij'} \leqslant C_l$ 或 $C_{ij'} \geqslant C_u$ 而停止”的情况下, SPRT 是根据同样观测个数进行的检验中具有最大检验力的假设检验, 即最优序贯检验。但是, 受制于现实情境下的疲劳效应、练习效应等因素的影响, 不可能要求被试持续作答直至满足上述条件(任赫, 陈平, 2021)。因此在 CCT 的实际使用中, 一般需要事先设定最大测验长度 J 。于是, 研究者在设计 CCT 终止规则时一般规定, 若在被试完成 J 道题目时测验仍未结束, 则通过下述准则对被试进行强制分类: 若 $C_{iJ} \leqslant C_0$, 则停止测验, 测验长度为 J, 并判断被试属于“未掌握”; 若 $C_{iJ} > C_0$, 则停止测验, 测验长度为 J, 并判断被试属于“掌握”。其中, $C_0 = (C_l + C_u)/2$ 。

将被试最终结束测验时实际作答的题目数记为 K, 对被试的分类判断结果记为 D (其中, $D = m$ 表示被试属于“掌握”, $D = n$ 表示被试属于“未掌握”), 则 SPRT 的判断准则可以概括如下,

$$\begin{cases} \text{停止测验}, K = j', D = n \\ \quad \text{若}\{j' < J, C_{ij'} \leqslant C_l\}\text{或}\{j' = J, C_{ij'} \leqslant C_0\} \\ \text{停止测验}, K = j', D = m \\ \quad \text{若}\{j' < J, C_{ij'} \geqslant C_u\}\text{或}\{j' = J, C_{ij'} > C_0\}. \\ \text{继续测验} \qquad\qquad\qquad\qquad \text{否则} \end{cases} \tag{3}$$

值得注意的是, 在 SPRT 中引入最大测验长度 J 虽然能够解决一些现实问题, 但是它违背 Wald-Wolfowitz 定理的前提假定, 导致 SPRT 不再是最优序贯检验。在现实测验中, 这不仅会增加测验长度和测验时间, 而且会提高题目曝光率。因此, 在维持 SPRT 的分类准确率基本不变的基础上, 尝试缩短测验长度可以减轻上述问题, 有助于 CCT 的应用。随机缩减技术(Finkelman, 2008; Huebner & Fina, 2015)正是一种尝试缩短测验长

度的方法。它的核心思想是：如果被试未来的作答不太可能会改变当前对被试的分类判断，那么此时便结束测验是合理的。

SCSPRT 规则就是一种将随机缩减技术与 SPRT 相结合的似然比终止规则。在完整保留公式(3)所定义的判断准则的基础上，SCSPRT 对原本需要继续作答的被试 i 进行预分类，并预测预分类结果能否保持。具体地说，SCSPRT 首先按照 SPRT 方法计算等式(2)所定义的对数似然比统计量 $C_{ij'}$，然后再计算“被试完成整份测验，即作答完 J 题后，得到的分类结果”与当前的预分类结果一致的概率。使用 $D_{j'}$ 表示在被试完成 j' 道题目后对被试的预分类；D_J 表示被试作答的题目数达到最大测验长度 J 时得到的分类结果。上述概率就可以表示为 $P(D_J = D_{j'}|C_{ij'})$，同样使用 C_0 作为判断准则。于是，对于预分类为“未掌握”的被试，$P(D_J = D_{j'}|C_{ij'}) = P(D_J = n|C_{ij'})$。

在公式(3)中，若 $j'<J$ 且 $C_l<C_{ij'}<C_u$，被试将继续进行作答。但是，SCSPRT 方法在 $j'<J$ 时，对公式(3)所定义的判断准则进行如下调整，

$$\begin{cases} \text{停止测验}, K = j', D = n \quad \text{若}\{C_{ij'} \leqslant C_l\}\text{或} \\ \quad \{C_l < C_{ij'} \leqslant C_0, P(D_J = n \mid C_{ij'}) \geqslant 1-\epsilon_1\} \\ \text{停止测验}, K = j', D = m \quad \text{若}\{C_{ij'} \geqslant C_u\}\text{或} \\ \quad \{C_u > C_{ij'} > C_0, P(D_J = m \mid C_{ij'}) \geqslant 1-\epsilon_2\}, \\ \text{继续测验} \qquad\qquad \text{否则} \end{cases} \tag{4}$$

其中，临界值 ϵ_1 与 ϵ_2 由测验开发者事先给定。以往的模拟研究表明：在保证一定分类精度的前提下，当 ϵ_1 与 ϵ_2 都取 0.05 时，SCSPRT 能大幅缩短测验长度(Finkelman, 2008, 2010)。$P(D_J = n|C_{ij'})$ 的具体计算详见 Finkelman (2008)与任赫和陈平(2021)。需要指出的是，在自适应选题的情境下，无法提前确定接下来选取的题目，这会给 $P(D_J = n|C_{ij'})$ 的计算带来一定困难。此时，可以选择一组“合适”的题目替代被试未来实际作答的题目。例如，若使用最大信息量选题策略，可以选择在被试“当前能力估计值”具有最大信息量的 $J-j'$ 道题作为替代题目。有研究者指出，如果使用替代题目，需适当减小 ϵ_1 和 ϵ_2 的取值(Finkelman, 2008)。

(2)多维的 SPRT 方法(C-SPRT、P-SPRT 与 M-SCSPRT)

在上述的 UCCT 中，通过事先确定的能力阈值 θ_0，可以很容易获得公式(1)中所需要的 θ_l 与 θ_u，并由此计算公式(2)所构造的 SPRT 统计量 $C_{ij'}$。但是在 MCCT 中，事先确定的只能是能力分界曲线或曲面，导致无法直接得到某个确定的阈值 θ_0。此外，即使获得 θ_0，多维空间中的 θ_0 在不同方向上可以构造任意多个 δ 邻域，因此如何选择可用于 $C_{ij'}$ 计算的 θ_l 和 θ_u 是另一个需要解决的问题。

C-SPRT 使用“约束在分界曲线上的能力估计值”作为能力分界点 θ_0 的近似($\hat{\theta}_0$)，并在该点处，沿分界曲线的法向量方向计算相应的 $\boldsymbol{\theta}_l$ 和 $\boldsymbol{\theta}_u$ 的近似($\hat{\boldsymbol{\theta}}_l$ 和 $\hat{\boldsymbol{\theta}}_u$)。具体地说，被试 i 完成 j' 道题目后，C-SPRT 算法首先在能力分界曲线或曲面($g(\boldsymbol{\theta}) = 0$)上计算能力参数的极大似然估计值，并将其作为阈值的近似，即

$$\hat{\boldsymbol{\theta}}_0 = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0}{\arg\max}\left[\log L(\boldsymbol{\theta}|\boldsymbol{Y}_{ij'})\right], \tag{5}$$

其中，$\boldsymbol{\Theta}_0 = \{\boldsymbol{\theta} : g(\boldsymbol{\theta}) = 0\}$。然后，在 $\hat{\boldsymbol{\theta}}_0$ 处沿 $g(\boldsymbol{\theta}) = 0$ 的法向量方向构造 δ 邻域。记 $\boldsymbol{\theta}_\delta$ 为该方向上的单位向量，可得到无差别区间的上、下限分别为 $\hat{\boldsymbol{\theta}}_u = \hat{\boldsymbol{\theta}}_0 + \delta\boldsymbol{\theta}_\delta$ 与 $\hat{\boldsymbol{\theta}}_l = \hat{\boldsymbol{\theta}}_0 - \delta\boldsymbol{\theta}_\delta$。最后，再按照 SPRT 构造似然比统计量(如公式(2))，就可以得到对数似然比统计量 $C_{ij'} = \log[\mathrm{LR}(\hat{\boldsymbol{\theta}}_u, \hat{\boldsymbol{\theta}}_l|\boldsymbol{Y}_{ij'})]$。

P-SPRT 与 C-SPRT 唯一的区别在于它采用“空间投影”而非“似然函数约束”的方法将分界曲线或曲面转换为可用于假设检验的分界点。具体地说，P-SPRT 将基于极大似然估计的被试能力估计值投影至能力分界曲线或曲面 $g(\boldsymbol{\theta}) = 0$ 上，并将投影点视作单维情境下能力阈值的近似，即

$$\hat{\boldsymbol{\theta}}_0 = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0}{\arg\min} \| \hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta} \|_2, \tag{6}$$

其中，$\hat{\boldsymbol{\theta}}_i$ 表示被试 i 的能力估计值，$\|\cdot\|_2$ 表示欧式空间中的距离。确定 $\hat{\boldsymbol{\theta}}_0$ 后，P-SPRT 与 C-SPRT 一样计算得到 $\hat{\boldsymbol{\theta}}_u$、$\hat{\boldsymbol{\theta}}_l$ 以及 $C_{ij'}$，并按照等式(3)所定义的准则对被试进行分类判断。

此外，与单维随机缩减方法类似，同样可以在多维似然比统计量的基础上融入随机缩减技术。M-SCSPRT 就是将多维情境下的 C-SPRT 与随机缩减相结合的终止规则。具体地说，与单维的 SCSPRT 类似，M-SCSPRT 使用 C-SPRT 的方法(等式(5))获得无差别区间的上下界，据此计算似然比统计量，并按照随机缩减技术计算 $P(D_J = D_{j'}|C_{ij'})$，进而根据公式(3)和(4)对被试进行分类。

2.1.2 多分类的 SPRT 方法

多分类情境是指测验要将被试划分到三个及以上的不同类别中。在此情境下，如果被试需要被分到 $S+1$ 个不同的类别之中，就需要定义 S 个能力分界点将不同被试区分开来。目前，在多分类情境下的终止规则研究仅限于 UCCT。这些研究在二分类方法的基础上，使用不同的思路在多个分界点处构造假设检验与检验统计量以完成对被试的分类。下面对单维多分类 CCT 中的 SPRT 方法进行介绍。

以上述的 $S+1$ 个类别为例，多分类的 SPRT 规则根据所确定的 S 个能力分界点，建立 S 个无差别区间以及与之对应的 S 个二分类 SPRT 检验。为便于理解，图 2 展示的是一个三分类问题的示意图。其中，θ_{1u} 和 θ_{1l} 分别表示能力分界点 θ_1 的无差别区间的上、下界，θ_{2u} 和 θ_{2l} 分别表示 θ_2 的无差别区间的上、下界。

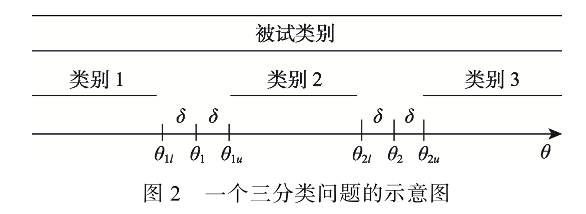


图 2 一个三分类问题的示意图

(1)Sobel-Wald 方法

Sobel 和 Wald (1949)所提出的多分类 SPRT 方法在每个能力分界点 θ_s 处，构建一组简单假设，即

$$\begin{aligned} H_{s0}&:\theta \leqslant \theta_{sl}=\theta_s-\delta \\ H_{s1}&:\theta \geqslant \theta_{su}=\theta_s+\delta, \end{aligned} \tag{7}$$

其中，θ_{su} 和 θ_{sl} 分别表示能力分界点 θ_s 所对应无差别区间的上、下界。基于公式(7)的假设检验，可以按照 2.1.1 中的 SPRT 构造似然比统计量，

$$C_{sij'}=\log\left[\frac{L(\theta_{su}|\boldsymbol{Y}_{ij'})}{L(\theta_{sl}|\boldsymbol{Y}_{ij'})}\right]. \tag{8}$$

由此，即可在每个 θ_s 处完成一组二分类的 SPRT 检验。Sobel-Wald 方法按照如下准则对被试进行分类判断：结合所有的 S 组检验，如果 H_{10} 被接受，就停止测验，测验长度为 j'，判断被试属于能力最低的类别，即类别 1；如果 H_{S1} 被接受，也停止测验，测验长度为 j'，判断被试属于能力最高的类别，即类别 $S+1$；如果 H_{s1} 和 $H_{(s+1)0}$ 被同时接受，同样停止测验，测验长度为 j'，判断被试属于类别 $s+1$；否则，就继续进行测验。如果测验达到最大长度 J，则停止测验，测验长度为 J，并根据被试 i 的能力极大似然估计值 $\hat{\theta}_i$ 和分界点的相对位置对其进行分类。该方法最早由 Eggen (1999)应用于 CCT，后来被 Thompson (2009)以及 van Groen 等人(2014)在三分类的情境下进行过评估。然而，Ghosh (1970)认为，在考虑更多的类别数时，Sobel-Wald 方法可能无法得出一个明确的分类判断。

(2)Armitage 方法

为解决 Sobel-Wald 方法可能无法得出结论的缺陷，Armitage (1950)提出一种比较所有可能的类别组合的 SPRT 方法。具体地说，对于 S 个能力分界点，就需要构造 $S(S+1)/2$ 组假设检验(Armitage, 1950; Seitz & Frey, 2013; Spray, 1993)。此时，任一组假设检验的原假设 H_p 与备择假设 H_q 分别表示考生属于类别 p 和($p<q\in\{2,\cdots,S+1\}$)，即

$$\begin{aligned} H_p&:\theta \leqslant \theta_{pl}=\theta_p-\delta \\ H_q&:\theta \geqslant \theta_{(q-1)u}=\theta_{q-1}+\delta. \end{aligned} \tag{9}$$

对应的检验统计量为，

$$C_{pqij'}=\log\left[\frac{L(\theta_{(q-1)u}|\boldsymbol{Y}_{ij'})}{L(\theta_{pl}|\boldsymbol{Y}_{ij'})}\right]. \tag{10}$$

Armitage 方法的分类准则为：如果所有包括假设 H_p 的检验都接受假设 H_p，则停止测验，测验长度为 j'，判断被试属于类别 p；否则，测试将继续进行，直到满足上述条件或达到最大测试长度为止。

需要说明的是，只有当 Sobel-Wald 方法无法给出准确的分类判断时，其与 Armitage 的方法才存在差异(Wang et al., 2021)。而在大多数情况下，这两种方法所得到的结果都一致，但是 Armitage 方法需要进行更多次检验。Wang 等人(2021)的研究中使用一个四分类问题为例，对其进行理论分析，感兴趣的读者可以参阅。也就是说，Sobel 和 Wald 方法在测验的分类准确率上应与 Armitage 方法相近，但在测验效率上应更胜一筹，这与已有研究的结果一致(Govindarajulu, 1987; Ghosh & Sen, 1991)。

2.2 广义似然比方法(GLR)

在 SPRT 中，最大测验长度的使用可能会降低分类准确率。为此，Bartroff 等人(2008)将 GLR 应用于 UCCT。之后，研究者又将随机缩减技术与 GLR

相结合, 提出随机缩减的 GLR 方法(Stochastically Curtailed GLR, SCGLR; Huebner & Fina, 2015)。另外, Nydick (2013)也将 GLR 方法推广到多维情境中, 提出多维的广义似然比方法(Multidimensional GLR, M-GLR)。

2.2.1　二分类的 GLR 方法

(1)单维的 GLR 方法(GLR 与 SCGLR)

不同于 SPRT 方法使用一组简单假设(即公式(1)), GLR 使用下述的一组复合假设对被试进行分类判断,

$$H_0:\theta\leqslant\theta_l \quad H_1:\theta\geqslant\theta_u. \tag{11}$$

由此, GLR 统计量 $C_{ij'}$ 是在"无差别区间"两侧各自的对数似然函数最大值之比, 即

$$C_{ij'}=\log\frac{\sup_{\theta_1\geqslant\theta_u}\left[L(\theta_1|\boldsymbol{Y}_{ij'})\right]}{\sup_{\theta_2\leqslant\theta_l}\left[L(\theta_2|\boldsymbol{Y}_{ij'})\right]}. \tag{12}$$

例如, 在图 1 中, 相比于 SPRT 使用在 θ_u 处的对数似然函数值(-9.68), GLR 使用在 θ_u 右侧的似然函数最大值(即 θ_l 处的 -6.60), 此时在 θ_l 左侧的似然函数最大值与 SPRT 时一致, 于是计算得到 $C_{ij'}=5.81$ 。在得到广义似然比统计量 $C_{ij'}$ 后, GLR 规则也按照公式(3)所定义的准则对被试进行分类判断。

此外, 也可以将随机缩减技术与 GLR 相结合, 得到随机缩减的 GLR 方法。与 SCSPRT 类似, SCGLR 是在 GLR 的基础上结合随机缩减技术而得到的。具体而言, 它使用与 GLR 方法相同的统计量 $C_{ij'}$ (如等式(12)所示), 然后根据随机缩减技术的要求计算 $P(D_J=D_{j'}|C_{ij'})$, 最后根据公式(3)和(4)对被试做出分类判断。

(2)多维的 GLR 方法

等式(12)所示的 GLR 统计量 $C_{ij'}$ 是在"无差别区间"两侧各自的对数似然函数最大值之比, 不再需要等式(2)中的 θ_u 和 θ_l 。因此, 在将 GLR 推广到 MCCT 时, 不再需要考虑如何进行"分界曲线或曲面"和分界点的转换的问题。M-GLR 统计量的定义为

$$C_{ij'}=\log\frac{\sup_{\boldsymbol{\theta}_1\in\Theta_m}\left[L(\boldsymbol{\theta}_1|\boldsymbol{Y}_{ij'})\right]}{\sup_{\boldsymbol{\theta}_2\in\Theta_n}\left[L(\boldsymbol{\theta}_2|\boldsymbol{Y}_{ij'})\right]}, \tag{13}$$

其中, Θ_m 表示多维空间中属于掌握类别的被试能力范围, Θ_n 表示多维空间中属于未掌握类别的被试能力范围。因此, 上式可以理解为对数似然函数在能力分界曲线或曲面两侧的最大值之比。与等式(12)所定义的单维 GLR 统计量相比, 等式(13)与其形式一致, 仅将似然函数求极大值的区域由两个单维的区间扩展到两个多维的空间。因此, 通过广义似然比的方式得到 $C_{ij'}$ 后, M-GLR 规则与 GLR 一样, 也是按照等式(3)的准则对被试进行分类。

2.2.2　多分类的 GLR 方法

回到 2.1.2 中 $S+1$ 个分类的问题, 针对所定义的 S 个无差别区间, 将由它们隔开的 $S+1$ 个不同类别的被试能力区间分别记为 $\Theta_1\equiv\{\theta\leqslant\theta_{1l}\},\cdots,\Theta_s\equiv\{\theta_{(s-1)u}\leqslant\theta\leqslant\theta_{sl}\},\cdots,\Theta_{S+1}\equiv\{\theta\geqslant\theta_{Su}\}$ 。按照 Wang 等人(2021)提出的多分类的 GLR 方法(multi-category GLR, mGLR), 得到如下的复合假设

$$H_s:\theta\in\Theta_s. \tag{14}$$

由此, Wang 等人(2021)指出可以根据序贯分析中的多假设 GLR 检验(Tartakovsky et al., 2014), 为上述复合假设构造如下的多分类 GLR 统计量

$$C_{ij'}^s=\log\left[\frac{\prod_{j=1}^{j'}L(\hat{\theta}_i\,|\,\boldsymbol{Y}_{ij'})}{\sup_{\theta\in\Theta_s}\prod_{j=1}^{j'}L(\theta\,|\,\boldsymbol{Y}_{ij'})}\right], \tag{15}$$

其中, 分子部分表示似然函数的极大值, 分母部分表示在能力区间 Θ_s 内似然函数的极大值。基于此, mGLR 方法定义如下的分类准则:

(1)当 $\hat{\theta}_i$ 不属于任何一个无差别区间时, 如果存在 s , 对所有 $t\neq s$, 有 $C_{ij'}^t\geqslant a_{st}$, 则停止测验, 测验长度为 j' , 判断被试属于类别 s 。这是因为当 H_s 为真时, 在 Θ_s 内的似然函数值会大于在其他无差别区域内的似然函数值, 从而使得 $C_{ij'}^s$ 的值较小而其他的 $C_{ij'}^t$ 的值较大。其中, a_{st} 是一个事先给定的值, 表示在 H_s 为真时接受 H_t 的概率。

(2)否则, 如果 $\hat{\theta}_i$ 落入由 θ_s 所定义的无差别区间, 则如等式(8)一样计算 $C_{sij'}$ 以决定将被试划分到类别 s 或 $s+1$ 或继续测验。

2.3　似然比规则简评

似然比检验的核心思想是比较有约束条件的似然函数的最大值与无约束条件的似然函数的最大值。如果两者之间的差异不大, 就可以认为对参数的约束有效; 反之, 则认为对参数的约束无效。基于此, 似然比规则在不同类别下, 建立符合该类别约束的似然函数, 并比较不同类别约束条

件的似然函数的最大值。如果某个类别的似然函数显著大于其他类别，就可以认为将考生划分到该类别是可信的，反之继续测验。由于似然比检验发展较为完备且具有良好的理论性质和检验效果，因此基于似然比检验的似然比规则是目前研究最为集中的一类 CCT 终止规则。已有研究也表明似然比规则还具有较好的稳健性，比如 Huang 等人(2000)认为即使题目参数没有得到准确标定，SPRT 方法也能获得较为准确的分类结果。但是，似然比规则也有一定的缺点，例如：(1) δ 的取值在很大程度上影响着 SPRT 方法的准确性与效率。尽管 δ 越大能使测验越快结束，但是大的 δ 会影响决策的精度。特别是对于多分类情境，如果 δ 过分大的话，不同的无差别区域很容易会出现重叠，从而使得我们很难去解释决策的结果。所以，δ 的取值范围是研究者需要注意的一个方面；(2)似然比规则在复杂测验情境(比如，多维和多分类)下的拓展比较复杂；(3)分界分数的选取具有较大的主观性。

3 贝叶斯决策理论规则

贝叶斯规则是另一类重要的 CCT 终止规则。不同于蕴含假设检验的规则，贝叶斯规则以贝叶斯决策理论为基础，通过定义后验概率与损失函数，就可以选择期望损失最小的决策以完成对被试的分类判断。其中，损失由错误决策所产生，具体可分为阈值损失和线性损失。目前为止，研究者对贝叶斯规则的研究基本仍限于 UCCT 情境。

3.1 阈值损失

(1)二分类的阈值损失规则

Lewis 和 Sheehan (1990)在二分类情境下提出一种阈值损失函数，也即用不同的常数来评估决策所有可能结果的损失。表 1 展示的是 Lewis 和 Sheehan (1990)的研究中，作答 j' 道题目后的阈值损失。

表 1 阶段 j' 时的二分类阈值损失函数

决策	$\theta=\theta_l$	$\theta=\theta_u$
被试属于“未掌握”	$j'l_c$	$l_{01}+j'l_c$
被试属于“掌握”	$l_{10}+j'l_c$	$j'l_c$

其中，l_c 表示被试作答一道题目的损失，以此控制测验效率(一般要求作答每道题目的损失是一样的)；l_{10} 为“将一位未掌握的考生划分到掌握类别”的损失，l_{01} 为“将一位掌握的考生划分到未掌握类别”的损失，以此控制测验精度。为简便起见，Lewis 和 Sheehan (1990)将各个测验阶段和各个决策的损失值都设置为相同。需要指出的是，正确分类所对应的损失 l_{00} 与 l_{11} 并未在表 1 中呈现。这是因为，这里假定正确分类的损失相同并且损失值非常小。表 1 所展示的损失函数是重新量尺化后的结果，l_{00} 与 l_{11} 在量尺转换后变为 0。此外，表 1 中并没有呈现“继续作答一道题目”的损失，这是因为继续作答的损失可以表示为与测验未来阶段中的分类决策(即掌握/未掌握)相关的损失的加权平均，权重等于得到相应决策的概率。

根据贝叶斯理论，被试 i 在作答 j' 道题后属于掌握类别的后验概率 $P_{m|j'}$ 可以如下式一般进行迭代计算，

$$P_{m|j'}=P(\theta=\theta_u|\boldsymbol{Y}_{i,j'})=\frac{P(\boldsymbol{Y}_{i,j'}|\theta_u)\cdot P_{m|j'-1}}{P(\boldsymbol{Y}_{i,j'}|\theta_u)\cdot P_{m|j'-1}+P(\boldsymbol{Y}_{i,j'}|\theta_l)\cdot P_{n|j'-1}},\quad(16)$$

其中，当 $j'=1$ 时，$P_{m|j'-1}$ 为被试为掌握类别的先验概率 P_m。并且，$P_{n|j'}=1-P_{m|j'}$。在被试作答题目数量为 j' 时，被试 i 被划分为掌握类别的期望损失(也称为风险函数)为，

$$\mathbb{E}_\theta[l(\theta,m)\,|\,\boldsymbol{Y}_{i,j'}]=j'\cdot l_c+l_{10}\cdot(1-P_{m|j'}),\quad(17)$$

其中，$l(\cdot,\cdot)$ 为损失函数。被试 i 此时被划分为未掌握类别的期望损失为，

$$\mathbb{E}_\theta[l(\theta,n)\,|\,\boldsymbol{Y}_{i,j'}]=j'\cdot l_c+l_{01}\cdot P_{m|j'}.\quad(18)$$

此外，被试 i 还可能被要求继续测验。而计算此时继续测验的期望损失就需要考虑在 $j'+1$ 时的所有可能决策的损失。为此，首先计算被试在第 $j'+1$ 道题目上的作答为 r 的概率，记为 $P_{r|j'}$。可以将 $P_{r|j'}$ 表示为 $P_{m|j'}$ 的函数，即，

$$P_{r|j'}=P(Y_{i,j'+1}=r\,|\,\boldsymbol{Y}_{i,j'})=P(Y_{i,j'+1}=r|\theta_l)P_{n|j'}+P(Y_{i,j'+1}=r|\theta_u)P_{m|j'},\quad(19)$$

其中，$P(Y_{i,j'+1}=r|\theta_l)$ 和 $P(Y_{i,j'+1}=r|\theta_u)$ 分别是“未掌握”与“掌握”的被试在第 $j'+1$ 题上作答为 r 的概率在整个题库水平上的平均值。

在贝叶斯规则中，使用最小化风险函数的方式给出决策。具体地说，在最大测验长度(即 $j=J$)时，由于必须对被试做出判断而不能继续要求被试作答，因此可以直接根据公式(17)和(18)给出此时的风险函数，并取使得风险函数最小的分类判

断作为决策。该决策可以表示为 $P_{m|J}$ 的函数，即

$$d_J(P_{m|J}) = \min\{\mathbb{E}_\theta[l(\theta,m)\,|\,\boldsymbol{Y}_{iJ}], \mathbb{E}_\theta[l(\theta,n)\,|\,\boldsymbol{Y}_{iJ}]\}. \quad (20)$$

在 $j<J$ 时，还需要考虑继续作答的损失。此时，根据上式就可以依次迭代，得到测验在达到最大长度之前继续作答的期望损失。比如，如图 3 所示，对于二级计分的题目，在 $j=J-1$ 时，被试分别以 $P_{0|J-1}$ 和 $P_{1|J-1}$ 的概率答错或答对下一题(第 J 题)。被试作答第 J 题后，由于达到最大测验长度，只需要做出分类决策而不需要继续作答，所以此时的风险函数就如同等式(20)。

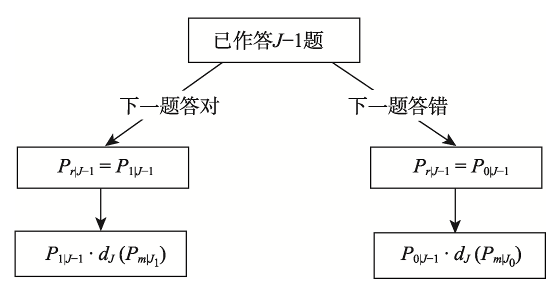


图 3　在第 $J-1$ 题时要求被试继续作答的损失(以二级计分题为例)

于是，在 $j=J-1$ 时要求被试继续作答的期望损失就可以用预期被试作答到第 J 题时的风险函数表示，即

$$\mathbb{E}_\theta[l(\theta,c)\,|\,\boldsymbol{Y}_{i,J-1}] = \sum_{r=0}^{1} P_{r|J-1}\cdot d_J(P_{m|J}) = P_{0|J-1}\cdot d_J(P_{m|J_0}) + P_{1|J-1}\cdot d_J(P_{m|J_1}), \quad (21)$$

其中，c 表示对考生的判断为需要继续做答，$P_{m|J_0}$ 和 $P_{m|J_1}$ 分别表示被试在第 J 道题上作答为 0 或 1 时被判断为掌握类别的后验概率，其计算按公式(19)进行。由此，在 $j=J-1$ 时的决策可记为，

$$d_{J-1}(P_{m|J-1}) = \min\{\mathbb{E}_\theta[l(\theta,m)\,|\,\boldsymbol{Y}_{i,J-1}], \mathbb{E}_\theta[l(\theta,n)\,|\,\boldsymbol{Y}_{i,J-1}], \mathbb{E}_\theta[l(\theta,c)\,|\,\boldsymbol{Y}_{i,J-1}]\}. \quad (22)$$

根据上式就可以对被试进行分类判断。具体地说，系统将选择使得期望损失最小的决定(将被试划分为掌握，未掌握或要求继续作答)，即

$$\begin{cases} \text{停止测验}, K=j', D=m & \text{若}\mathbb{E}_\theta[l(\theta,m)\,|\,\boldsymbol{Y}_{i,J-1}]\text{最小} \\ \text{停止测验}, K=j', D=n & \text{若}\mathbb{E}_\theta[l(\theta,n)\,|\,\boldsymbol{Y}_{i,J-1}]\text{最小}. \\ \text{继续测验} & \text{若}\mathbb{E}_\theta[l(\theta,c)\,|\,\boldsymbol{Y}_{i,J-1}]\text{最小} \end{cases} \quad (23)$$

以此类推，就可以得到在 $j=j'$ 时被试继续作答的期望损失，并选择使得期望损失最小的决定完成对被试的判断。

(2)多分类的阈值损失规则

对于贝叶斯规则而言，从二分类到多分类的推广比较简单。对于一个三分类的 UCCT，只需要将表 1 中的阈值损失函数替换为表 2 中内容，再选择最小的损失即可完成对被试的分类判断(Vos, 1999)。

表 2　阶段 j' 的三分类阈值损失函数

决策	$\theta \leqslant \theta_{1l}$	$\theta_{1u}<\theta<\theta_{2l}$	$\theta \geqslant \theta_{2u}$
被试属于“类别 1”	$j'l_c$	$l_{12}+j'l_c$	$l_{13}+j'l_c$
被试属于“类别 2”	$l_{21}+j'l_c$	$j'l_c$	$l_{23}+j'l_c$
被试属于“类别 3”	$l_{31}+j'l_c$	$l_{32}+j'l_c$	$j'l_c$

3.2　线性损失

表 1 中的阈值损失函数具有一个明显的缺点：它假定对于不同能力值的被试的损失是恒定的，而不考虑这些被试能力值与分界分数的距离。但事实上，能力值离分界分数更远的被试被错误分类所造成的损失往往更严重。此外，阈值损失函数的值也不是连续变化的，这在很多情况下也不符合现实。因此，一种更合理的假设是：损失函数是关于能力与分界分数间距离的连续增函数(van der Linden & Mellenbergh, 1977; van der Linden & Vos, 1996; Vos, 1997a, 1997b)。

(1)二分类的线性损失规则

Van der Linden 和 Mellenbergh (1977)在二分类情境下，提出一种线性损失函数，如表 3 所示。可以发现，相比于阈值损失，线性损失使得决策成本可以随“能力值 θ 离分界分数 θ_0 的距离”的变化而线性变化。

表 3　阶段 j' 的二分类线性损失函数

决策	$\theta=\theta_l$	$\theta=\theta_u$
被试属于“未掌握”	$j'l_c$	$b_1(\theta_0-\theta)+j'l_c$
被试属于“掌握”	$b_2(\theta_0-\theta)+j'l_c$	$j'l_c$

其中，斜率 b_1 与 b_2 是由有经验的专家确定。在给定损失函数后，就可以按照后验概率得到损失最小的决策，从而完成对被试的分类。

(2)多分类的线性损失规则

与阈值损失函数类似，在多分类情境下，只需要将表 3 中的线性损失函数替换成表 4 中的内容即可得到一种三分类的线性损失函数(Vos, 1999)。

表 4 阶段 j' 的三分类线性损失函数

决策	$\theta \leqslant \theta_{1l}$	$\theta_{1u} < \theta < \theta_{2l}$	$\theta \geqslant \theta_{2u}$
被试属于"类别 1"	$j'l_c$	$b_1(\theta-\theta_1)+j'l_c$	$b_1(\theta-\theta_1)+j'l_c$
被试属于"类别 2"	$b_2(\lvert\theta-\theta^*\rvert-\theta')+j'l_c$	$j'l_c$	$b_2(\lvert\theta-\theta^*\rvert-\theta')+j'l_c$
被试属于"类别 3"	$b_3(\theta_2-\theta)+j'l_c$	$b_3(\theta_2-\theta)+j'l_c$	$j'l_c$

其中，$\theta^*=(\theta_2+\theta_1)/2, \theta'=(\theta_2-\theta_1)/2$ 。

3.3 贝叶斯规则简评

贝叶斯规则所提供的思路与似然比规则的完全不同。似然比规则是通过构造似然比统计量进行假设检验，贝叶斯方法则是通过作答更新被试能力的后验分布，并使用后验概率计算损失函数值，从而基于贝叶斯决策论完成对被试的判断。

需要指出的是，在贝叶斯规则中，有无数种可能的损失函数，没有哪一种损失函数一定是最好的。这一特点既是贝叶斯规则最大的优点，也是其饱受诟病的一点。支持者认为这使得该方法能够考虑多样的损失函数，具有更大的灵活性；但是，反对者认为损失函数的选择具有一定程度的任意性。在使用该方法之前，研究者需要考虑清楚如何客观、科学地选择需要的损失函数。

4 置信区间规则

除似然比规则和贝叶斯规则外，CCT 终止规则中还有一种是 ACI 方法。ACI 方法通过比较分界分数与"被试能力估计值的置信区间"的相对位置，来完成对被试的分类判断。

4.1 置信区间规则介绍

目前，对于这种方法的研究较少且集中在二分类的 UCCT 中。值得注意的是，ACI 中所涉及到的被试能力估计，既可以使用极大似然估计也可以使用贝叶斯估计。具体而言，如果使用极大似然估计，则通过测量标准误(Standard Error of Measurement, SEM)构造置信区间；如果使用贝叶斯估计，则使用贝叶斯后验方差的平方根构造置信区间。分类测验过程中，不断更新的被试 i 的能力估计值的置信区间可以表示为，

$$\hat{\theta}_i - z_\epsilon \times \delta_{error} < \theta < \hat{\theta}_i + z_\epsilon \times \delta_{error}, \qquad (24)$$

其中，z_ϵ 为$(1-\epsilon)$的置信区间所对应的标准正态分布分位数，$\epsilon=\alpha+\beta$ 为两类错误率之和，δ_{error} 表示对能力的极大似然估计中的 SEM 或贝叶斯估计中后验方差的平方根。例如，如果设置第 I 类、第 II 类错误率均为 0.025，那么 ϵ 为 0.05，这时 z_ϵ 等于 1.96。在极大似然估计中，SEM 根据被试 i 的所有已作答题目的 Fisher 信息量计算，即

$$SEM = \frac{1}{\sqrt{\Sigma_{j=1}^{j'} I_{ij}}}, \qquad (25)$$

其中，I_{ij} 表示题目 j 为被试 i 提供的 Fisher 信息量，对 j' 道题目的信息量求和即得到该被试在已作答的 j' 道题目上的总信息量。Thompson (2011)的研究指出，有两种方式可以实现等式(25)的计算：一是理论最大值的 SEM；二是观察分数的 SEM。根据被试已作答题目所组成的测验，理论最大值的 SEM 是在被试能力所有可能取值的范围内每隔一定步长(比如，在[−3,3]的区间内每隔 0.01)计算一个 SEM，并取其最大值；观察分数的 SEM 则是在被试的能力估计值处，计算 SEM。简小珠和陈平(2020)指出，在大多数研究中都使用观察分数的 SEM 进行计算。

得到置信区间 $[\hat{\theta}_i - z_\epsilon \times SEM, \hat{\theta}_i + z_\epsilon \times SEM]$ 后，ACI 方法的分类准则如下：如果分界分数低于该区间的下界(即 $\hat{\theta}_i - z_\epsilon \times SEM$)，那么停止测验，测验长度为 j'，并判断被试属于"掌握"；如果分界分数高于该区间的上界(即 $\hat{\theta}_i + z_\epsilon \times SEM$)，那么停止测验，测验长度为 j'，并判断被试属于"未掌握"；否则就继续进行测验。

4.2 置信区间规则简评

在某种程度上，可以认为 ACI 方法将被试的分类问题转化为被试的能力估计问题。这样做的好处是使得对被试的分类变得非常直观、简洁。但是，这种方法的稳健性相对较差。因为使用该方法需要有足够大的标定题库作为前提，否则就可能会导致较高的错误率。同时，Eggen 和 Straetmans (2000)以及 Thompson (2009)的研究都表明：该方法所需的测验长度一般高于似然比规则。

5 三类终止规则的综合分析

5.1 三类终止规则的构造思路与优缺点分析

综上所述，三类终止规则各有优缺点。其中，似然比规则基于似然比检验，具有较好的理论性

质，大多数测验情境下最为准确、高效，相关研究也较多。但是，由于需要定义无差别区间大小和第 I、第 II 类错误率，引入了主观因素的影响，并且该方法在多维、多分类等复杂测验情境下的拓展难度较大。已有的多分类 SPRT 终止规则(Sobel-Wald 方法与 Armitage 方法)是对多个能力分界点独立进行假设检验，因此会隐含多重比较的问题，即实际的第 I 和第 II 类错误率远大于设定标准。尽管已有研究者留意到这一点(Wang, 2019; Wang et al., 2021)，但由于第 I 和第 II 类错误率的变化并不是影响 SPRT 规则的分类准确性的主要因素，所以较少有研究对其进行校正。

贝叶斯规则通过后验概率与损失函数，完成对被试的分类判断。该方法无需事先给定第 I 和第 II 类错误率，它以更全局的角度动态优化决策，从测验最后的阶段向前倒推，因此每一步的损失判断都能考虑到整个测验过程。损失函数的多样性使得该方法的形式非常灵活，也使得该方法很容易就可以被应用于不同的测验情境中。但是，该方法也存在一定问题：(1)当结合 IRT 模型时，从后向前的损失函数计算量会变得十分巨大，不利于该方法的实施；(2)正如公式(19)所示，已有的贝叶斯方法在计算“下一道题得到特定作答的概率，即 $P(Y_{i,j'+1}=r|\theta_l)$ 或 $P(Y_{i,j'+1}=r|\theta_u)$”时使用的是在题库所有题目上的概率的平均值(即认为每道题目在下一阶段具有同等地位)，这显然不符合自适应的特性；(3)损失函数形式的灵活会不可避免地导致使用者在损失函数的选择上产生疑问，也可能会在实际应用中产生由于损失函数选取不恰当而导致的误差。

ACI 方法直接将分界分数与能力估计值的置信区间进行比较，无需划定无差别区间，并且计算简单且计算量小，是三种方法中最直接的一类方法。但是，这种方法的稳健性较差，测验效率也相对较低。表 5 是对上述各种方法的总结。

5.2　三类终止规则的适用情境

需要指出的是，CCT 是一个非常复杂的测验系统。终止规则的优劣还会受到 CCT 中其他部分(比如，心理测量模型、题库结构和选题策略)以及被试能力分布等因素的影响，三类终止规则在不同的测验情境下各占鳌头。因此，实践者在选择终止规则时需要综合考虑 CCT 的各个部分以明确

表 5　CCT 终止规则的总结

核心原理	类别数	维度数	终止规则	构造思路
似然比规则				
序贯似然比	二分类	单维	SPRT	在分界点处构造一组简单假设及对应的序贯似然比统计量
			SCSPRT	在 SPRT 的基础上结合随机缩减技术
		多维	C-SPRT	通过似然函数约束转化为 SPRT
			P-SPRT	通过欧氏空间投影转化为 SPRT
			M-SCSPRT	在 C-SPRT 的基础上结合随机缩减技术
	多分类	单维	Sobel-Wald 方法	在每个分类点处进行一次 SPRT
			Armitage 方法	为所有可能的类别组合进行 SPRT
广义似然比	二分类	单维	GLR	在分界点处构造一组复杂假设及对应的广义似然比统计量
			SCGLR	在 GLR 的基础上结合随机缩减技术
		多维	M-GLR	将 GLR 中的能力区间转化为多维能力空间
	多分类	单维	mGLR	对被试属于每个类别构造一组复杂假设及对应的广义似然比统计量
贝叶斯规则				
阈值损失	二分类	单维	Lewis-Sheehan 方法	确定每种决策所对应的损失
	多分类		Vos 方法	确定每种决策所对应的损失
线性损失	二分类		Linden-Mellenbergh 方法	确定每种决策所对应的损失，并考虑能力估计值与分界点的距离
	多分类		Vos 方法	确定每种决策所对应的损失，并考虑能力估计值与分界点的距离
置信区间规则				
置信区间	二分类	单维	ACI	比较能力估计值的置信区间与分界点的相对位置

三类终止规则的适用情境。另外，还需要注意相应情境下可能面临的现实问题。

对于似然比规则，想要准确且快速做出决策的关键在于最大程度地区分不同类别被试的似然函数值，而这通常和选题策略密切相关。举例而言，在 UCCT 中，两种常见的选题策略是基于能力估计值的最大信息量选题方法(estimate-based maximum Fisher information)和基于分界分数的最大信息量选题方法(cutscore-based maximum Fisher information)。因此，当选题策略为后者时，所选的题目能够为假设检验提供更多的信息，因此似然比规则在基于分界分数的最大信息量选题方法下的效率最高。但是由于基于分界分数会因为固定点选题而导致题目高曝光的问题，所以似然比规则更适用于低风险的测验，而且要求题库中大部分题目在分界分数处具有高信息量。此外，由于 GLR 考虑无差别区间两侧的所有对数似然函数值(不仅着眼于上、下界两个点)，所以相比于 SPRT，GLR 在基于当前能力估计选题时也能保持一定的效率。

对于贝叶斯规则，高效分类的关键在于最大程度地区分不同决策损失的差异。由于不同决策损失函数的计算同样基于 θ_l 和 θ_u，所以在基于分界分数的选题方法下会有更好的表现。同样，考虑到题目曝光率的问题，贝叶斯方法更适用于低风险的测验。另外，由于贝叶斯方法能够针对不同的决策损失进行控制，所以适用于需要降低特定类型决策损失的测验。

对于置信区间规则，保障决策效率的关键在于不断地减小能力估计标准误。因此，ACI 方法在基于能力估计值的最大信息量选题方法下的效率最高，该选题策略可以减小置信区间的大小。此外，根据不同被试的能力，ACI 规则能够为不同被试呈现不同的题目，在一定程度上能降低高信息量题目的曝光率，所以它可以用于高风险的测验，相应地需要题库中的题目在不同能力位置具有高信息量。但是 Tian (2018)在控制分类准确性一致的前提下，采用基于能力估计值的选题方法，比较单维二分类的似然比规则和置信区间规则。结果发现：当被试能力分布远离分界分数时，ACI 规则的效率要高于似然比规则；但是在被试能力分布靠近分界分数时，ACI 规则效率低于 GLR 方法。这意味着 ACI 规则的表现还会受到被试能力分布与分界分数相对位置的影响，因此更适用于要求高通过率或低通过率的测验。

6 未来研究方向及应用

6.1 CCT 终止规则的未来研究方向

本文对多种测验情境下的CCT终止规则进行系统梳理与述评。目前，对 CCT 终止规则的研究已经比较丰富，但仍有一些地方有待完善。未来研究方向主要表现在以下四方面：

(1)完善基于贝叶斯的终止规则。构建 CCT 终止规则的思路主要有三个角度，即似然比方法、贝叶斯方法和置信区间方法。基于似然比方法的终止规则已经得到充分的发展，但如前所述，以贝叶斯方法为基础的终止规则仍然较少。未来，研究者可以考虑基于贝叶斯方法对前人研究进行完善。例如，在现实测验情景中，除考虑决策的准确率和测验长度之外，还需要满足其他非统计约束(如：内容均衡，即让试卷充分涵盖所要考察的知识模块)。由于贝叶斯损失函数具有灵活性，研究者可以考虑将各种非统计约束纳入终止规则的考虑范围。此外，正如 5.1 部分所言，目前贝叶斯方法没有利用已有的信息对被试即将作答的下一道题进行预测，未来研究可以借鉴似然比方法中随机缩减的思想来构造一组“合适”的题目替代被试未来实际作答的题目。最后，研究者还可以对损失函数中损失值的选取如何影响测验结果进行讨论。

(2)开发多维多分类的 CCT 终止规则。多维或多分类的 CCT 终止规则是近期的一个研究热点，但尚未有研究者探究同时满足多维、多分类要求的 CCT 终止规则。在现实应用中，许多测验不仅要同时考察被试在多个维度上的潜在特质，而且也需要将被试分到多于两个的类别中。例如，教育工作者希望将学生的数学成就水平划分为基础、熟练和高级三个类别(比如，美国国家进步教育评估 NAEP)；而数学测验也往往同时考察学生的算术、阅读和问题解决能力等，呈现出多维的能力结构(Reckase, 2009)。这就对构建多维、多分类的 CCT 终止规则提出迫切需求。

(3)开发融合作答时间(Response Time, RT；詹沛达 等, 2020)的 CCT 终止规则。近几年来，心理测量学的研究重点大都放在如何同时衡量多个维度的潜在特质，以向被试提供更详细、更完善的

反馈。但是这些研究大多只考虑被试的作答信息, 而很少使用行为信息。在 CCT 测验中, 有一类很容易获得的行为信息, 即被试作答所用的时间。Sie 等人(2015)尝试构建融入 RT 的 CCT, 他们的研究结果表明：融入 RT 后, 测验在分类精度轻微提高的同时还能够减少平均测验时间。但是, Sie 等人(2015)的研究主要集中在限制被试作答时间, 而未考虑更普遍的限制测验长度的情况。未来, 研究者可以在上述研究的基础上进一步展开探索, 开发新的结合 RT 的 CCT 终止规则, 在保持判断准确率的基础上缩短测验长度, 而不仅仅是控制测验时间。另外, 可以考虑如何利用作答时间提高分类决策的精度, 进而间接提高测验效率(Man et al., 2019; 詹沛达, 2019)。

(4)开发结合机器学习算法的 CCT 终止规则。目前的三类终止规则均为基于心理测量模型的方法, 模型的正确设定和前提假设的满足对结果有重要的影响, 然而实践中的数据往往掺杂着各式各样的噪音。机器学习是近年来各个领域研究的热点, 其中许多算法都是用来解决分类问题, 这与 CCT 的目的相一致。Gonzalez (2021)认为, 相比于比较“通过各种模型估计得到的被试能力”与“黄金标准”来获得被试的类别, 机器学习算法通过被试的作答就能直接预测被试属于某个类别的概率, 避免模型不拟合等引起的误差。Zheng 等人(2020)基于机器学习算法中的决策树方法, 开发出一个短的基于树的自适应分类测验。未来, 研究者可以考虑使用其他的分类算法(比如, 逻辑斯蒂克回归、支持向量机以及随机森林等方法)完成自适应分类测验。

6.2　CCT 终止规则的应用

CCT 测验主要包含两种类型：合格性测验与临床医学问卷。在为不同类型的测验制定终止规则时, 应充分考虑测验的考生群体、试题特点以及决策影响。

在合格性测试中, 通过设置不同难度的试题, 将考生划分到不同能力水平, 根据考生的等级水平, 来决定其从业资格、学业进度或升学。许多职业资格考试都属于这类测验, 比如教师资格考试、司法考试和执业医师资格考试等; 此外, 还有一些学业水平考试也属于合格性测验, 比如大学英语四、六级考试、计算机二级考试以及初中学业水平测试等。对于此类测验, 往往每年均有数量庞大的考生群体, 具有充足的测验经费和考生样本, 相应地能够建立起一定规模的题库, 并在一定程度上能保障题目参数的稳定估计, 使得合格性测验具有运用三类终止规则的潜力。但是, 似然比规则与贝叶斯规则的原理较为复杂, 且正如 5.2 部分所言, 这些方法在实践中伴随着题目曝光率过高的问题。因此, 在现有的合格性测验尤其是高风险的合格性测验中, 鲜有这两类方法的应用。与上述两种规则的困境形成对比的是, 置信区间规则原理简明易懂、分类结果清晰, 更能为大众和教育工作者所理解, 更具有推广性, 在现实中就显得更加可行。比如, 美国联合委员会注册护士执照考试(the National Council Licensure Examination for Registered/Practical Nurse, NCLEX-RN)就使用 ACI 规则来决定测验何时终止。

在临床医学问卷中, 通过评价患者在不同指标上的轻重程度或近期的心理生理状态, 将患者划分到不同症状水平, 来为其后续的治疗和诊断提供依据。比如, 汉密尔顿抑郁量表(Hamilton Rating Scale for Depression, HRSD)和创伤后应激障碍量表(Posttraumatic Stress Disorder Checklist, PCL)。对于此类测验, 被试群体往往很小, 且问卷的题项并不具有一般意义上的难度。更重要的是, 假阴性(false negative)的分类结果所带来的代价不可忽视。因此, 考虑到相比于另外两类终止规则, 贝叶斯终止规则能够对各种分类损失有更精细的控制, 在临床医学问卷中更为适用。目前, 终止规则在临床医学问卷中的应用目的主要为：在保证决策准确基础上缩短已有问卷的长度, 使得诊断过程更高效, 比如利用机器学习模型或随机缩减技术进一步缩减问卷长度(Gonzalez, 2021; Smits et al., 2016)。还需要注意的是, 临床问卷以往直接使用观测分数与诊断临界值相比较, 而已有的终止规则主要基于潜在特质进行计算。但随着 IRT 研究的推进, 越来越多的研究者使用 IRT 模型对临床问卷建模, 比如 Li 等人(2019)将等级反应模型(Graded Response Model, GRM)应用于病人健康问卷(the Patient Health Questionnaire, PHQ)。因此, 相比于 Smits 等人(2016)使用基于观测分数的 CCT 并选择随机缩减的倒计时法(countdown method)作为终止规则, 贝叶斯规则或许既能够缩短测验长度, 又能在每一步中严格控制诊断的损失。

参考文献

陈平. (2016). 两种新的计算机化自适应测验在线标定方法. *心理学报, 48*(9), 1184–1198.

简小珠, 陈平. (2020). 计算机化分类测验的特点与发展述评. *考试研究,* (6), 77–89.

康春花, 辛涛. (2010). 测验理论的新发展:多维项目反应理论. *心理科学进展, 18*(3), 530–536.

任赫, 陈平. (2021). 两种新的多维计算机化分类测验终止规则. *心理学报*, *53*(9), 1044–1058.

詹沛达. (2019). 计算机化多维测验中作答时间和作答精度数据的联合分析. *心理科学*, (1), 170–178.

詹沛达, Hong Jiao, Kaiwen Man. (2020). 多维对数正态作答时间模型：对潜在加工速度多维性的探究. *心理学报, 52*, 1132–1142.

Armitage, P. (1950). Sequential analysis with more than two alternative hypotheses, and its relation to discriminant function analysis. *Journal of the Royal Statistical Society, Series B, 12*(1), 137–144.

Bartroff, J., Finkelman, M., & Lai, T. L. (2008). Modern sequential analysis and its applications to computerized adaptive testing. *Psychometrika, 73*(3), 473–486.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23*(3), 249–261.

Eggen, T. J. H. M., & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement, 60*(5), 713–734.

Ferguson, R. L. (1969). *Computer-assisted criterion-referenced measurement* (Working Paper No. 41). Pittsburgh, PA: University of Pittsburgh, Learning and Research Development Center.

Finkelman, M. (2003). *An adaptation of stochastic curtailment to truncate Wald's SPRT in computerized adaptive testing* (CSE Report 606). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Finkelman, M. (2008). On using stochastic curtailment to shorten the SPRT in sequential mastery testing. *Journal of Educational and Behavioral Statistics, 33*(4), 442–463.

Finkelman, M. (2010). Variations on stochastic curtailment in sequential mastery testing. *Applied Psychological Measurement, 34*(1), 27–45.

Finkelman, M., He, Y., Kim, W., & Lai, A. M. (2011). Stochastic curtailment of health questionnaires: A method to reduce respondent burden. *Statistics in Medicine, 30*(16), 1989–2004.

Ghosh, B. K. (1970). *Sequential tests of statistical hypotheses*. Reading, MA: Addison-Wesley.

Ghosh, B. K., & Sen, P. K. (1991). *Handbook of sequential analysis*. New York, NY: Marcel Dekker.

Gonzalez, O. (2021). Psychometric and machine learning approaches for diagnostic assessment and tests of individual classification. *Psychological Methods, 26*(2), 236–254.

Govindarajulu, Z. (1987). *The sequential statistical analysis of hypothesis, testing, point and interval estimation, and decision theory (American series in mathematical and management sciences).* Columbus, OH: American Sciences Press, Inc.

Huang, C.-Y., Kalohn, J. C., Lin, C.-J., & Spray, J. (2000). *Estimating item parameters from classical indices for item pool development with a computerized classification test* (Research Report 2000-4). Iowa City, IA: ACT, Inc.

Huebner, A. R., & Fina, A. D. (2015). The stochastically curtailed generalized likelihood ratio: A new termination criterion for variable-length computerized classification tests. *Behavior Research Methods, 47*(2), 549–561.

Kingsbury, G. G., & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 257–283). New York, NY: Academic Press.

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14*(4), 367–386.

Li, C., Moore, S. C., Smith, J., Bauermeister, S., & Gallacher, J. (2019). The costs of negative affect attributable to alcohol consumption in later life: A within-between random longitudinal econometric model using UK Biobank. *PLOS ONE, 14*(2), Article e0211357. https://doi.org/10.1371/journal.pone.0211357

Man, K., Harring, J. R., Jiao, H., & Zhan, P. (2019). Joint modeling of compensatory multidimensional item responses and response times. *Applied Psychological Measurement*, *43*(8), 639–654.

Nydick, S. (2013). *Multidimensional mastery testing with CAT* (Unpublished doctoral dissertation). University of Minnesota.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 237–257). New York, NY: Academic Press.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Seitz, N.-N., & Frey, A. (2013). The sequential probability ratio test for multidimensional adaptive testing with between-item multidimensionality. *Psychological Test and Assessment Modeling, 55*(1), 105–123.

Sie, H., Finkelman, M. D., Riley, B., & Smits, N. (2015).

Utilizing response times in computerized classification testing. *Applied Psychological Measurement, 39*(5), 389–405.

Smits, N., & Finkelman, M. D. (2013). A comparison of computerized classification testing and computerized adaptive testing in clinical psychology. *Journal of Computerized Adaptive Testing, 1*, 19–37.

Smits, N., Finkelman, M. D., & Kelderman, H. (2016). Stochastic curtailment of questionnaires for three-level classification: Shortening the CES-D for assessing low, moderate, and high risk of depression. *Applied Psychological Measurement, 40*(1), 22–36.

Sobel, M., & Wald, A. (1949). A sequential decision procedure for choosing one of three hypotheses concerning the unknown mean of a normal distribution. *Annals of Mathematical Statistics, 20*(4), 502–522.

Spray, J. A. (1993). *Multiple-category classification using a sequential probability ratio test* (ACT Research Report Series, No. 93-7). Iowa City, IA: Americn College Testing.

Spray, J. A., & Reckase, M. D. (1996). Comparison of SPRT and sequential Bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics, 21*(4), 405–414.

Tartakovsky, A., Nikiforov, I., & Basseville, M. (2014). *Sequential analysis: Hypothesis testing and changepoint detection*. Boca Raton, FL: Chapman and Hall/CRC.

Thompson, N. A. (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement, 69*(5), 778–793.

Thompson, N. A. (2011). Termination criteria for computerized classification testing. *Practical Assessment, Research, & Evaluation, 16*(4), 1–7.

Tian, C. (2018). *Comparison of four stopping rules in computerized adaptive testing and examination of their application to on-the-fly multistage testing* (Unpublished master's thesis). University of Illinois.

van der Linden, W. J., & Mellenbergh, G. J. (1977). Optimal cutting scores using a linear loss function. *Applied Psychological Measurement, 1*(4), 593–599.

van der Linden, W. J., & Vos, H. J. (1996). A compensatory approach to optimal selection with mastery scores. *Psychometrika, 61*, 155–172.

van Groen, M. M., Eggen, T. J. H, M., & Veldkamp, B. P. (2014). Item selection methods based on multiple objective approaches for classifying respondents into multiple levels. *Applied Psychological Measurement, 38*(3), 187–200.

Vos, H. J. (1997a). Simultaneous optimization of quota-restricted selection decisions with mastery scores. *British Journal of Mathematical and Statistical Psychology, 50*(1), 105–125.

Vos, H. J. (1997b). A simultaneous approach to optimizing treatment assignments with mastery scores. *Multivariate Behavioral Research, 32*(4), 403–433.

Vos, H. J. (1999). Applications of Bayesian decision theory to sequential mastery testing. *Journal of Educational and Behavioral Statistics, 24*(3), 271–292.

Wald, A. (1947). *Sequential analysis*. New York, NY: John Wiley.

Wald, A., & Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics, 19*, 326–339.

Wang, C., Chen, P., & Huebner, A. (2021). Stopping rules for multi-category computerized classification testing. *British Journal of Mathematical and Statistial Psychology, 74*(2), 184–202.

Wang, Z. (2019). *Grid multi-classification adaptive classification testing with multidimensional polytomous items* (Unpublished doctoral dissertation). University of Minnesota.

Zhan, P., Jiao, H., Man, K., Wang, W.-C., & He, K. (2021). Variable speed across dimensions of ability in the joint model for responses and response times. *Frontiers in psychology, 12*, Article 469196. https://doi.org/10.3389/fpsyg.2021.469196

Zheng, Y., Cheon, H., & Katz, C. M. (2020). Using machine learning methods to develop a short tree-based adaptive classification test: Case study with a high-dimensional item pool and imbalanced data. *Applied Psychological Measurement, 44*(7–8), 499–514. https://doi.org/10.1177/0146621620931198

Types, characteristics and application of termination rules in computerized classification testing

REN He, HUANG Yingshi, CHEN Ping

(*Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing Normal University, Beijing* 100875, *China*)

Abstract: Computerized classification testing (CCT) has been widely used in eligibility testing and clinical psychology for its efficiency in classifying participants. As an essential part of CCT, the termination rule determines when the test is to be stopped and what category the participants are ultimately classified into, directly affecting the test efficiency and classification accuracy. According to the theoretical basis of the termination rules, existing rules can be roughly divided into the likelihood ratio, Bayesian decision theory, and confidence interval rules. And their core ideas are constructing hypothesis tests, designing loss functions, and comparing the relative positions of confidence intervals, respectively. Based on these ideas, in different test situations, CCT termination rules have various specific forms. Future research can further extend Bayesian rules, construct rules for multidimensional and multicategory CCT, integrate process data into termination rules, and build rules under the framework of machine learning. In addition, from the perspective of practical requirement, all three types of rules have the potential to be applied in eligibility tests, while the Bayesian rules are optimal to clinical questionnaires.

Key words: computerized classification testing, termination rule, likelihood radio, stochastic curtailment, Bayesian decision theory